

Exposé zum Thema

Robustheit in der Regressionsanalyse: Ein systematischer Vergleich von OLS, Ridge, Lasso und bayesianischen Modellen unter Verletzung der Modellannahmen

Masterarbeit

Fachbereich: Data Science / Statistik

Vorgelegt von:

[Vorname, Nachname]

05.08.2026

Inhaltsverzeichnis

Problemstellung & Relevanz.....	1
Forschungsfrage und Forschungsziel.....	1
Forschungsziel und Erkenntnisinteresse.....	2
Forschungsstand und theoretische Grundlagen.....	2
Theoretischer Rahmen und Begriffsklärungen.....	3
Forschungskonzept/Methodisches Vorgehen.....	3
Eingrenzung und Umfang.....	4
Vorläufige Gliederung.....	4
Vorläufiges Literaturverzeichnis.....	6

Problemstellung & Relevanz

Die klassische Kleinste-Quadrate-Schätzung (OLS) gilt unter den Gauss-Markov-Annahmen als bester linearer erwartungstreuer Schätzer (BLUE). In der statistischen Praxis führen jedoch Ausreißer, Multikollinearität und Heteroskedastizität häufig zu einer massiven Verzerrung der Ergebnisse und instabilen Koeffizienten, da OLS einen Breakdown-Point von Null besitzt (Rousseeuw, 1997). Während frequentistische Regularisierungsverfahren wie Ridge und Lasso Lösungen für Multikollinearität bieten (Hoerl & Kennard, 1970; Zou, 2006), versprechen bayesianische Ansätze durch die Wahl robuster Prioren eine stabilere Modellierung von Unsicherheit (Park & Casella, 2008).

Trotz der theoretischen Vorteile robuster und bayesianischer Verfahren fehlt oft ein direkter, simulationsbasierter Vergleich der Performance dieser Modellklassen unter simultaner Verletzung mehrerer Annahmen (Maronna et al., 2006).

In der Data Science ist die Wahl des richtigen Modells bei unsauberen Datensätzen kritisch für die Vorhersagegüte und die Validität statistischer Inferenz, insbesondere wenn keine realen Parameter für die Validierung vorliegen.

Forschungsfrage und Forschungsziel

Inwieweit weisen regularisierte frequentistische und bayesianische Regressionsmodelle eine überlegene Robustheit gegenüber der klassischen OLS-Regression auf, wenn Ausreißer, Heteroskedastizität und Multikollinearität in verschiedenen Intensitäten auftreten?

Unterfragen:

- Welchen Einfluss hat der Anteil von Ausreißern auf den Mean Squared Error (MSE) der Koeffizientenschätzungen bei OLS im Vergleich zu Ridge und Lasso?
- Inwiefern stabilisieren bayesianische Prioren (z.B. Laplace-Prior) die Schätzung bei hoher Multikollinearität im Vergleich zum frequentistischen Lasso?
- Wie verhält sich die Vorhersagegenauigkeit der Modelle bei Vorliegen von Heteroskedastizität?

Hypothesen:

- **H1:** Regularisierte Verfahren (Ridge, Lasso) zeigen bei hoher Multikollinearität eine geringere Varianz der Schätzer als OLS, erkauft durch einen geringen Bias (Hoerl & Kennard, 1970).
- **H2:** Bayesianische Regressionsmodelle mit schwerfälligen Priorern erzielen bei hohen Ausreißeranteilen stabilere Ergebnisse als klassische OLS-Modelle (West, 1984).

Forschungsziel und Erkenntnisinteresse

Bestehende Studien vergleichen oft nur zwei Modellklassen (z.B. OLS vs. Lasso) oder konzentrieren sich auf eine einzelne Annahmeverletzung (z.B. nur Multikollinearität). Ein systematischer 'Stress-Test', der OLS, Ridge, Lasso und bayesianische Modelle unter kombiniertem Einfluss von Ausreißern, Heteroskedastizität und Multikollinearität vergleicht, ist in der aktuellen Literatur unterrepräsentiert.

Diese Arbeit schließt diese Lücke durch ein einheitliches Simulationsdesign, das die Grenzen der Robustheit für alle vier Modellklassen simultan aufzeigt.

Die Ergebnisse liefern eine Entscheidungsgrundlage für die Modellwahl in Data-Science-Pipelines und dienen als Ausgangspunkt für die Entwicklung hybrider Schätzer, die sowohl bayesianische als auch regularisierende Vorteile vereinen.

Forschungsstand und theoretische Grundlagen

Die Literatur etabliert M-Schätzer und LTS als Standard für den Umgang mit Ausreißern (Hampel, 2001; Rousseeuw, 1997). Maronna (2011) zeigt, dass MM-Schätzer hohe Effizienz mit Robustheit verbinden können.

Es besteht Konsens, dass Lasso die Variablenselektion optimiert, während Ridge bei Multikollinearität die Varianz reduziert (Tibshirani, 1996; Hoerl & Kennard, 1970). Zou (2006) erweiterte dies um das Adaptive Lasso zur Korrektur von Selektionsbias.

Das Bayesian Lasso (Park & Casella, 2008) integriert die L1-Strafe in einen Wahrscheinlichkeitsrahmen und bietet natürliche Unsicherheitsintervalle. West (1984) und Mitchell & Beauchamp (1988) wiesen die Vorteile von Spike-and-Slab-Priorern für die Variablenselektion nach.

Es wird debattiert, ob bayesianische Modelle bei sehr großen Datensätzen (Large p , Small n) den rechentechnisch effizienteren frequentistischen Modellen überlegen sind (Tew et al., 2023).

Theoretischer Rahmen und Begriffsklärungen

Die Arbeit basiert auf dem Gauss-Markov-Theorem und dessen modernen Erweiterungen (Hansen, 2022). Ergänzend wird die statistische Entscheidungstheorie (Fourdrinier et al., 2018) herangezogen, um den Bias-Varianz-Tradeoff bei Shrinkage-Schätzern zu bewerten.

Zentrale Konzepte:

- **Breakdown Point:** Das Maß für die Robustheit eines Schätzers gegenüber kontaminierten Daten (Hampel, 2001).
- **Shrinkage:** Die bewusste Einführung von Bias zur Reduktion der Varianz (Hoerl & Kennard, 1970).
- **Posterior-Verteilung:** Die Verteilung der Parameter nach Einbezug von Daten und Prior-Information (Park & Casella, 2008).
- **Heteroskedastizität:** Nicht-konstante Varianz der Fehlerterme, die OLS ineffizient macht (Lakshmi et al., 2021).

Forschungskonzept/Methodisches Vorgehen

Die Arbeit liefert einen quantitativen Vergleich der Schätzgüte (MSE, Bias, Varianz) und der Vorhersagepräzision (RMSE) über verschiedene Simulationsszenarien hinweg.

Argumentationslinie:

- **Schritt 1:** Theoretische Herleitung der Schwächen von OLS bei Annahmeverletzungen basierend auf der Literatur (Rousseeuw, 1997).
- **Schritt 2:** Durchführung von Simulationen, die zeigen, dass Ridge und Lasso bei Multikollinearität stabilere Koeffizienten liefern als OLS.
- **Schritt 3:** Nachweis, dass das Bayesian Lasso durch die Posterior-Verteilung eine robustere Unsicherheitsquantifizierung bei Ausreißern ermöglicht.

Es wird erwartet, dass bayesianische Modelle insbesondere in Szenarien mit geringen Stichproben und hohen Ausreißerraten den frequentistischen Ansätzen überlegen sind.

Eingrenzung und Umfang

Die Arbeit konzentriert sich auf lineare Regressionsmodelle für stetige Zielvariablen. Untersucht werden OLS, Ridge-Regression (L2), Lasso-Regression (L1) und das Bayesian Lasso. Die Evaluation erfolgt ausschließlich über Monte-Carlo-Simulationen mit kontrollierten Parametern (n , p , SNR, Korrelationsstruktur).

Nicht-lineare Modelle (z.B. Random Forests, neuronale Netze) werden nicht berücksichtigt. Ebenso werden keine realen Datensätze verwendet, um die Vergleichbarkeit gegenüber den 'wahren' Parametern der Simulation zu gewährleisten. Zeitreihenanalytische Besonderheiten (Autokorrelation) werden zugunsten der klassischen Querschnittsannahmen vernachlässigt.

Vorläufige Gliederung

1. Einleitung (ca. 5 Seiten)

- Problemstellung und Zielsetzung
- Aufbau der Arbeit

2. Theoretische Grundlagen der Regressionsanalyse (ca. 12 Seiten)

2.1 Das klassische lineare Modell und OLS

2.2 Gauss-Markov-Theorem und BLUE-Eigenschaft

2.3 Verletzungen der Modellannahmen: Konsequenzen für OLS

3. Regularisierungsverfahren in der frequentistischen Statistik (ca. 10 Seiten)

3.1 Ridge-Regression: L2-Penalty und Multikollinearität

3.2 Lasso-Regression: L1-Penalty und Variablenselektion

3.3 Der Bias-Varianz-Tradeoff bei Shrinkage-Methoden

4. Bayesianische Regressionsanalyse (ca. 12 Seiten)

4.1 Bayesianische Inferenz und Prior-Spezifikationen

4.2 Das Bayesian Lasso: Hierarchische Modellierung

4.3 MCMC-Verfahren zur Posterior-Schätzung

- 5. Konzepte der robusten Statistik (ca. 8 Seiten)
 - 5.1 Breakdown Point und Einflussfunktionen
 - 5.2 Robuste Verlustfunktionen und M-Schätzer
- 6. Methodik der Simulationsstudie (ca. 10 Seiten)
 - 6.1 Design der Monte-Carlo-Simulation
 - 6.2 Definition der Störszenarien und Metriken
 - 6.3 Implementierung in R (brms, glmnet)
- 7. Ergebnisse und Diskussion (ca. 18 Seiten)
 - 7.1 Performanzvergleich unter Multikollinearität
 - 7.2 Robustheit gegenüber Ausreißern und Heteroskedastizität
 - 7.3 Diskussion der praktischen Implikationen
- 8. Fazit und Ausblick (ca. 5 Seiten)

Vorläufiges Literaturverzeichnis

- Adnan, N., Ahmad, M. H., & Adnan, R. (2006). A Comparative Study On Some Methods For Handling Multicollinearity Problems. *MATEMATIKA*, 22, Article 2.
- Alma, Ö. G. (2011). Comparison of Robust Regression Methods in Linear Regression. *International Journal of Contemporary Mathematical Sciences*, 6, Article 9.
- Babacan, S. D., Molina, R., & Katsaggelos, A. K. (2010). Bayesian Compressive Sensing Using Laplace Priors. *IEEE Transactions on Image Processing*, 19, Article 1. <https://doi.org/10.1109/TIP.2009.2032894>
- Basu, A., Ghosh, A., Jaenada, M., & Pardo, L. (2023). Robust adaptive Lasso in high-dimensional logistic regression [Research report]. arXiv. <https://arxiv.org/abs/2109.03028>
- Brown, S. H. (2009). Multiple Linear Regression Analysis: A Matrix Approach with MATLAB. *Alabama Journal of Mathematics*, 1–3.
- Bürkner, P.-C. (2024, September 23). brms: Bayesian Regression Models using “Stan”. RDocumentation. <https://www.rdocumentation.org/packages/brms/versions/2.22.0>
- Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, 11, 1–11. <https://doi.org/10.1002/wics.1460>
- Chaibub Neto, E., Bare, J. C., & Margolin, A. A. (2014). Simulation Studies as Designed Experiments: The Comparison of Penalized Regression Models in the “Large p, Small n” Setting. *PLoS ONE*, 9, Article 10. <https://doi.org/10.1371/journal.pone.0107957>
- Cook, R. D., & Weisberg, S. (1982). Diagnostics for Heteroscedasticity in Regression (Technical Report 405). University of Minnesota.
- Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91, Article 434.

- Croux, C., Dhaene, G., & Hoorelbeke, D. (2003). Robust Standard Errors for Robust Estimators [Discussion paper]. Katholieke Universiteit Leuven.
- Dar, Y., Muthukumar, V., & Baraniuk, R. G. (2021). A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning [Preprint]. arXiv. <https://arxiv.org/abs/2109.02355>
- Djalil, I., & Terzic, S. (2021). Violation of the Assumption of Homoscedasticity and Detection of Heteroscedasticity. *Decision Making: Applications in Management and Engineering*, 4, Article 1. <https://doi.org/10.31181/dmame2104001d>
- Fourdrinier, D., Strawderman, W. E., & Wells, M. T. (2018). Shrinkage Estimation. Springer. <https://doi.org/10.1007/978-3-030-02185-6>
- Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., & Yang, J. (2023). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models [R package]. CRAN. <https://doi.org/10.32614/CRAN.package.glmnet>
- Fu, Y.-H., Fillion, K. B., Bodnar, L. M., Brooks, M. M., Platt, R. W., Himes, K. P., & Naimi, A. I. (2020). Visualization tool of variable selection in bias–variance tradeoff for inverse probability weights. *Annals of Epidemiology*, 41, 95–99. <https://doi.org/10.1016/j.annepidem.2019.12.004>
- Geyer, C. J. (1994). On the Asymptotics of Constrained M-Estimation. *The Annals of Statistics*, 22, Article 4.
- Geyer, C. J. (2006). Breakdown Point Theory Notes [Class notes]. University of Minnesota.
- Geyer, C. J. (2011). Introduction to Markov Chain Monte Carlo. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 3–48). CRC Press. <https://doi.org/10.1201/b10905-2>
- Greven, S. (2023). Flexible Dichteregressionsmethoden (513634041). Deutsche Forschungsgemeinschaft. GEPRIS - Geförderte Projekte der DFG. <https://gepris.dfg.de/gepris/projekt/513634041>

- Halliwel, L. J. (2015). The Gauss-Markov Theorem: Beyond the BLUE. *Casualty Actuarial Society E-Forum*, 1–38.
- Hampel, F. (2001). Robust statistics: A brief introduction and overview (Research Report 94). Seminar für Statistik, ETH Zürich.
<https://www.research-collection.ethz.ch/bitstreams/11b6700a-8cf0-44e4-aeed-563b4fcb2efc/download>
- Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics*, 190, 115–132.
<https://doi.org/10.1016/j.jeconom.2015.09.003>
- Hansen, B. E. (2022). A Modern Gauss–Markov Theorem. *Econometrica*, 90, Article 3. <https://doi.org/10.3982/ECTA19255>
- Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2020). Best Subset, Forward Stepwise, or Lasso? Analysis and Recommendations Based on Extensive Comparisons. 1–19.
<https://www.stat.cmu.edu/~ryantibs/papers/bestsubset.pdf>
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12, Article 1.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.
<https://doi.org/10.1007/978-1-4614-7138-7>
- Joshi, C., Ruggeri, F., & Wilson, S. P. (2017). Prior robustness for Bayesian implementation of the fault tree analysis. *IEEE Transactions on Reliability*, 1–15.
- Kim, J., & Li, J. C.-H. (2023). Which Robust Regression Technique Is Appropriate Under Violated Assumptions? A Simulation Study. *Methodology*, 19, Article 4. <https://doi.org/10.5964/meth.8285>
- Kim, M. G. (2017). A cautionary note on the use of Cook’s distance. *Communications for Statistical Applications and Methods*, 24, Article 3.
<https://doi.org/10.5351/CSAM.2017.24.3.317>
- Kipruto, E., & Sauerbrei, W. (2022). Comparison of variable selection procedures and investigation of the role of shrinkage in linear

- regression-protocol of a simulation study in low-dimensional data. PLOS ONE, 17, Article 10. <https://doi.org/10.1371/journal.pone.0271240>
- Kowalskie, A. (2025). The Impact of Outliers on Linear Regression Models: Detection and Correction Strategies. OTS Canadian Journal, 4, Article 6. <https://doi.org/10.58840/fzbcv732>
- Kumar, N. K. (2023). Autocorrelation and Heteroscedasticity in Regression Analysis. Journal of Business and Social Sciences, 5, Article 1. <https://doi.org/10.3126/jbss.v5i1.72442>
- Lakshmi, K., Mahaboob, B., Rajaiah, M., & Narayana, C. (2021). Ordinary least squares estimation of parameters of linear model. Journal of Mathematical and Computational Science, 11, Article 2. <https://doi.org/10.28919/jmcs/5454>
- Lee, G. Y. (2021). Regression Shrinkage and Selection for Actuarial Models. Variance, 14, Article 2.
- Leng, C., Tran, M. N., & Nott, D. (2010). Bayesian Adaptive Lasso. arXiv. <https://arxiv.org/abs/1009.2300>
- Li, J., Wu, R., Lu, I.-T., & Ren, D. (2023). Bayesian Linear Regression with Cauchy Prior and Its Application in Sparse MIMO Radar. IEEE Transactions on Aerospace and Electronic Systems, 1–23. <https://arxiv.org/abs/2307.11233>
- Loris, I., Douma, H., Nolet, G., Daubechies, I., & Regone, C. (2010). Nonlinear regularization techniques for seismic tomography. Geophysical Journal International, 181, Article 2. <https://doi.org/10.1111/j.1365-246X.2010.04533.x>
- Maronna, R. A. (2011). Robust ridge regression for high-dimensional data. Technometrics, 53, Article 2. <https://doi.org/10.1198/TECH.2011.09131>
- Maronna, R. A., Martin, D. R., & Yohai, V. J. (2006). Robust Statistics: Theory and Methods (Vol. 835). Wiley. <https://books.google.nl/books?id=iFVjQgAACAAJ>
- Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. Journal of Statistical Software, 42, Article 9.

- Martín, N., & Pardo, L. (2008). On the Asymptotic Distribution of Cook's distance in Logistic Regression Models [Research report]. Complutense University of Madrid.
- Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83, Article 404.
- Moore, R. C., & DeNero, J. (2011). L1 and L2 Regularization for Multiclass Hinge Loss Models. *Machine Learning in Speech and Language Processing (MLSLP)*, 1–5. <http://www.isca-speech.org/archive>
- Müller, C. (2019). Robuste Statistik [Vorlesungsskript]. Technische Universität Dortmund. https://stabi.statistik.tu-dortmund.de/storages/stabi-statistik/r/Lehre/SoSe_19/Robuste_Statistik/RobStat.pdf
- Newey, W. K. (1995). Convergence Rates and Asymptotic Normality for Series Estimators (Working Paper 95-13). Massachusetts Institute of Technology. <http://www.archive.org/details/convergencerates00newe2>
- Ogutu, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, 6. <https://doi.org/10.1186/1753-6561-6-S2-S10>
- Orenti, A., Zolin, A., Marubini, E., Antonelli, P., Ambrogi, F., & Cesana, B. M. (2024). Robust Regression as a Sensible Alternative to the Weighted Ordinary Least Squares Regression in case of Heteroskedasticity. A Tutorial. *Epidemiology Biostatistics and Public Health*, 19, Article 1. <https://doi.org/10.54103/2282-0930/26484>
- Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103, Article 482. <https://doi.org/10.1198/016214508000000337>
- Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92, Article 437.
- Ranstan, J., & Cook, J. A. (2016). LASSO regression. *BJS*.

- Reid, S., Tibshirani, R., & Friedman, J. (2014). A Study of Error Variance Estimation in Lasso Regression. arXiv, 1–30.
<https://arxiv.org/abs/1311.5274>
- Reiß, M. (2011). Nichtparametrische Statistik: Skript zur Vorlesung im Wintersemester 2010/11 [Vorlesungsskript]. Humboldt-Universität zu Berlin.
- Rousseeuw, P. J. (1997). Introduction to Positive-Breakdown Methods. In G. S. Maddala & C. R. Rao (Eds.), *Handbook of Statistics* (Vol. 15, pp. 101–121). Elsevier.
- Schreiber-Gregory, D. N. (2018). Ridge Regression and Multicollinearity: An In-Depth Review. SAS Global Forum 2018 Proceedings, Article 2825-2018.
<https://support.sas.com/resources/papers/proceedings18/2825-2018.pdf>
- Simpson, D. G., Carroll, R. J., & Ruppert, D. (1985). M-estimation for discrete data: Asymptotic distribution theory and implications [Technical report]. University of North Carolina.
- Stahel, W. (2017). Statistische Regressionsmodelle: Teil I: Lineare Regression [Vorlesungsunterlagen]. Seminar für Statistik, ETH Zürich.
- Startz, R. (2012). Bayesian Heteroskedasticity-Robust Standard Errors [Working paper]. University of California, Santa Barbara.
<https://escholarship.org/uc/item/69c4x8m9>
- Stepniak, C. (2010). From Equivalent Linear Equations to Gauss-Markov Theorem. *Journal of Inequalities and Applications*, 2010, Article 259672.
<https://doi.org/10.1155/2010/259672>
- Tew, S. Y., Boley, M., & Schmidt, D. F. (2023). Bayes beats Cross Validation: Fast and Accurate Ridge Regression via Expectation Maximization. 37th Conference on Neural Information Processing Systems (NeurIPS 2023), 1–20. <https://github.com/marioboley/fastridge.git>
- University of Colorado Boulder. (2018). An Introduction to Bayesian Linear Regression. In APPM 5720: Bayesian Computation. University of Colorado Boulder.
https://www.colorado.edu/amath/sites/default/files/attached-files/bayes_linreg.pdf

- van den Bergh, D., Clyde, M. A., Gupta, A. R. K. N., de Jong, T., Gronau, Q. F., Marsman, M., Ly, A., & Wagenmakers, E.-J. (2021). A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behavior Research Methods*, 53, 2351–2371. <https://doi.org/10.3758/s13428-021-01552-2>
- van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*, 25, 143–154. <https://doi.org/10.3758/s13423-016-1015-8>
- van Wieringen, W. N. (2023). Lecture notes on ridge regression (0.60 ed.). arXiv. <https://arxiv.org/abs/1509.09169>
- Welsh, A. H. (1989). On M-Processes and M-Estimation. *The Annals of Statistics*, 17, Article 1. <https://doi.org/10.1214/aos/1176347022>
- West, M. (1984). Outlier Models and Prior Distributions in Bayesian Linear Regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46, Article 3. <https://doi.org/10.1111/j.2517-6161.1984.tb01311.x>
- Yu, C., Yao, W., & Bai, X. (2014). Robust Linear Regression: A Review and Comparison [Research report]. Kansas State University. <https://arxiv.org/abs/1404.6274>
- Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101, Article 476. <https://doi.org/10.1198/016214506000000735>
- Zuo, Y. (2019). Robustness of the deepest projection regression functional [Preprint]. arXiv. <https://arxiv.org/abs/1806.09611>